DOCUMENT RESUME

ED 187 665 SP 015 614

AUTHOR Capie, William; And Others
TITLE Selected Investigations of the Reliability of the Teacher Performance Assessment Instruments (A Working Paper).
INSTITUTION Georgia Univ., Athens. Coll. of Education.
SPONS AGENCY Georgia State Dept. of Education, Atlanta.
PUB DATE 79
NOTE 63p.

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Data Collection; Evaluation Criteria; *Evaluation Methods; *Teacher Evaluation; *Test Reliability; Test Validity

ABSTRACT

A study investigating the degree of agreement between teacher evaluators using a standard set of evaluation criteria and evaluation methods examines the test reliability of the Classroom Procedures Assessment Instrument, the Interpersonal Skills Assessment Instrument, and the Teaching Plans and Materials Assessment Instrument. Two statistical indexes resulted from the study. One index illustrates the amount of exact agreement between raters; the second indicates the amount of near agreement. The indexes, along with other statistical information, are compiled for six different cases of investigation: 1) interrater agreement using the TPAI in training situations (such as with beginning teachers); 2) internal consistency of ratings using the TPAI; 3) interrater agreement using the TPAI classrooms; 4) interrater agreement using the TPAI in training situations (student teachers); 5) internal consistency of ratings using the TPAI (student teachers); and, 6) interrater agreement using the TPAI in classrooms (student teacher). (LH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED187665

Georgia Department of Education
Student/Beginning Teacher Evaluation Study

Selected Investigations of the Reliability of the Teacher Performance Assessment Instruments

(A Working Paper)

William Capie Ed.D.
Chad D Ellett Ph.D.
Charles E Johnson Ed.D.

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

1979

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
A. Meughon
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Teacher Assessment Project
College of Education, The University of Georgia
Athens, Georgia 30602

SP 015614

Table of Contents

I. Interrater Agreement Using the TPAI in Training Situations (Beginning Teacher) . . 1

II. Internal Consistency of Ratings Using the TPAI (Beginning Teacher) 11

III. Interrater Agreement Using the TPAI in Classrooms (Beginning Teacher) 18

IV. Interrater Agreement Using the TPAI in Training Situations (Student Teacher) . . . 28

V. Internal Consistency of Ratings Using the TPAI (Student Teacher) 36

VI. Interrater Agreement Using the TPAI in Classrooms (Student Teacher) 40

Interrater Agreement Using the TPAI in Training Situations (Beginning Teacher Edition)

The Problem

To what extent do raters assign the same TPAI descriptor rating when they observe the same teaching performance?

Procedures

This study was conducted as part of the training activities for data coll[illegible]-trainers in Athens, during the summer of 1978. There were six sessi[illegible]hose participants represented a variety of experience levels and educational positions.

Participants viewed videotapes and used the Classroom Procedures Assessment Instrument (CP) and the Interpersonal Skills Assessment Instrument (IS) to rate several classroom lessons as practice and one as a final evlauation. They also evaluated sets of lesson plans using the Teaching Plans and Materials Assessment Instrument (TPM).

Two similar types of data analyses were completed. Each resulted in an index (proportion) of agreement among the raters. The first index was Exact Agreement, which amounted to choosing the same or identical rating for an indicator. In other words, the portion of raters choosing the mode was determined. The second index was Near Agreement, which amounted to selecting either the mode or the next most common adjacent scale point. The proportions of near and exact agreement were calculated for each indicator in each training session. The mean rate of agreement acorss the six lessons was also computed.

Instances in which all or most of the raters failed or abstained to rate an item were still treated as numerical ratings when determining Exact Agreement. Thus, if all raters agreed that an indicator was not

rateable then 100 percent Exact and Near agreement were noted. Instances where some raters failed or abstained to rate an item were recorded as "misses" when Near Agreement was computed. Samples of these procedures are included in Table 1.

Table 1

Sample Ratings and Proportion of Agreement

Teacher	Indicator	Number of Ratings For Each Rating Point 1	2	3	4	5	X	Proportion of Exact-Agreement	Proportion of Near-Agreement
A	CP 1			14	13			.52	1.00
B	TPM 5			4	13	12		.45	.86
C	IS 3			14		14		.50	.50
D	TPM 11			15	3	9		.56	.67
E	CP 22				3	24		.89	.89

Results

The percentages of Near and Exact Agreement for each indicator of the TPM are summarized in Table 2. The mean level of Exact Agreement ranged from .40 (indicator 6) to .69 (indicator 9) with an overall mean of .52. The mean level of Near Agreement ranged from .56 (indicator 6) to .93 (indicator 5) with a grand mean of .77. When considering the TPM as a whole the mean near agreement was approximately 50 percent higher than the mean exact agreement.

Table 3 summarizes percentages of Near and Exact Agreement among the raters using the Classroom Procedures (CP) instrument. The mean level of Exact Agreement ranged from .52 (indicator 18) to .91 (indicator 21) with the grand mean of .62. Near Agreement varied from .82 (indicator 3) to .95 (indicator 21) with an overall mean of .88. The mean levels of Exact Agreement for the four lessons fell between .58 (3rd Grade Science) to .64 (9th Grade Consumer Economics and 8th Grade Language Arts).

Table 4 contains a summary of the rates of agreement for the Interpersonal Skills Assessment Instrument (IS). The mean levels of Exact Agreement ranged from .43 (indicator 6) to .80 (indicator 8). Near Agreement ranged from .72 to .94 (indicators 6 and 7, respectively). Near Agreement was approximately 42 percent higher than Exact Agreement. The mean rates of Exact Agreement over the four lessons were fairly consistent (.64, .54, .60, .59).

Table 2

Interrater Agreement Using Standard Interviews with the TPM

Indicator	8th Grade Language Arts Plans (n=67) Exact Agreement	8th Grade Language Arts Plans (n=67) Near Agreement	3rd Grade Science Plans (n=112) Exact Agreement	3rd Grade Science Plans (n=112) Near Agreement	Mean (n=179) Near Agreement	Mean (n=179) Near Agreement
1	.28	.52	.49	.81	.41	.70
2	.46	.81	.40	.70	.42	.74
3	.46	.75	.51	.88	.49	.83
4	.76	.85	.35	.64	.50	.72
5	.52	.88	.61	.96	.58	.93
6	.48	.88	.36	.36	.40	.56
7	.49	.79	.54	.73	.52	.75
8	.58	.90	.55	.81	.56	.84
9	.66	.82	.70	.82	.69	.82
10	.78	.91	.54	.84	.63	.87
11	.57	.76	.31	.59	.41	.65
12	.75	.90	.41	.80	.54	.84
13	.48	.66	.54	.87	.52	.79
14	.81	.85	.61	.76	.68	.79
15	.66	.87	.38	.60	.48	.70
Mean=	.58	.81	.49	.74	.52	.77

Table 3

Interrater Agreement Using
the CP with Videotaped Lessons

Indicator	9th Grade Consumer Economics (n=41)		Int. Language Arts (n=27)		8th Grade Language Arts (n=68)		3rd Grade Science (68)		Mean	
	Exact	Near	Exact	Near	Exact	Near	Exact	Mean	Exact	Mean
1.	.66	.88	.59	.85	.59	.91	.43	.69	.57	.83
2.	.54	.85	.52	.93	.60	.94	.47	.79	.53	.88
3.	.44	.78	.44	.81	.62	.85	.62	.82	.53	.82
4.	.49	.87	.52	.89	.75	.90	.50	.68	.57	.84
5.	.46	.85	.89	1.00	.76	.85	.60	.85	.68	.89
6.	.76	.98	.78	.93	.47	.88	.56	.91	.64	.93
7.	.93	.95	.41	.70	.79	.91	.81	.87	.74	.86
8.	.66	.95	.81	1.00	.72	.97	.53	.81	.68	.93
9.	.66	1.00	.70	.89	.65	.85	.46	.71	.62	.86
10.	.56	.85	.67	.89	.47	.85	.47	.74	.54	.83
11.	.54	.93	.56	.85	.54	.79	.44	.76	.53	.83
12.	.71	1.00	.41	.78	.49	.98	.66	.81	.57	.89
13.	.49	.83	.59	.81	.54	.78	.53	.93	.54	.84
14.	.51	.83	.67	1.00	.56	.94	.47	.90	.55	.92
15.	.66	.87	.52	.85	.57	.91	.69	.87	.61	.88
16.	.71	.90	.78	.96	.76	.90	.49	.85	.69	.90
17.	.66	.93	.52	.81	.50	.90	.43	.81	.53	.86
18.	.61	.90	.48	.78	.50	.91	.49	.91	.52	.88
19.	.51	.90	.48	.89	.78	.91	.79	.94	.64	.91
20.	.71	.90	.48	.78	.50	.99	.69	.93	.60	.90
21.	.93	.95	1.00	1.00	.96	.99	.76	.85	.91	.95
22.	.87	.93	.85	.93	.85	.91	.78	.85	.84	.91
Mean	.64	.90	.62	.88	.64	.90	.58	.83	.62	.88

TABLE 4
Interrater Agreement Using The IS
With Videotaped Lessons

Indicator	9th Consumer Education n=41		Int. Language Arts n=30		8th Languate Arts n=71		3rd Science n=72		Total n=214	
	Exact	Near	Exact	Near	Exact	Near	Exact	Near	Exact	Near
1	.71	1.00	.87	1.00	.42	.69	.46	.93	.62	.91
2	.44	.63	.57	.83	.42	.75	.68	.97	.53	.80
3	.61	.90	.50	1.00	.68	.87	.52	.93	.58	.93
4	.51	.90	.47	.87	.56	.89	.53	.92	.53	.90
5	.46	.78	.47	.77	.45	.75	.40	.64	.45	.74
6	.49	.76	.50	.87	.41	.65	.32	.60	.43	.72
7	.96	1.00	.47	.90	.58	.93	.64	.94	.66	.94
8	.85	.90	.53	.67	.94	.99	.88	.94	.80	.88
9	.76	.85	.50	.70	.93	.97	.89	.93	.77	.86
Mean	.64	.86	.54	.85	.60	.83	.59	.87	.60	.85

Discussion

The results of these studies are useful for identifying indicators that possess ambiguities in various rating contexts. The data are presented graphically in Figure 1. The TPAI indicators are listed on the abscissa and the rates of agreement are plotted on the ordinate. Indicators 1, 2, 6, and 11 posed difficulty in the attainment of Exact Agreement. Indicators 1 and 2 are part of a set that is difficult to rate because the "appropriateness" of activities and/or materials for learners must be judged. In training, these decisions may be required of raters who are working with an unfamiliar grade or subject area. Indicators 11 and 6 seemed to have been particularly difficult in the Third Grade Science lesson. This lesson may have had bits of ambiguous interview information. The mean level of Exact Agreement was slightly above fifty percent for this instrument in the two situations rated.

The mean level of agreement on the CP instrument was higher (sixty-two percent) than that for the TPM. This figure may have been inflated somewhat because indicators 21 and 22 are often left unrated and thusly easily agreed upon. Several of the indicators (2, 4, 10, 11, 13, 14, 17, and 18) have rates of agreement below the mean for the instrument, but none is so low as to suggest a problem.

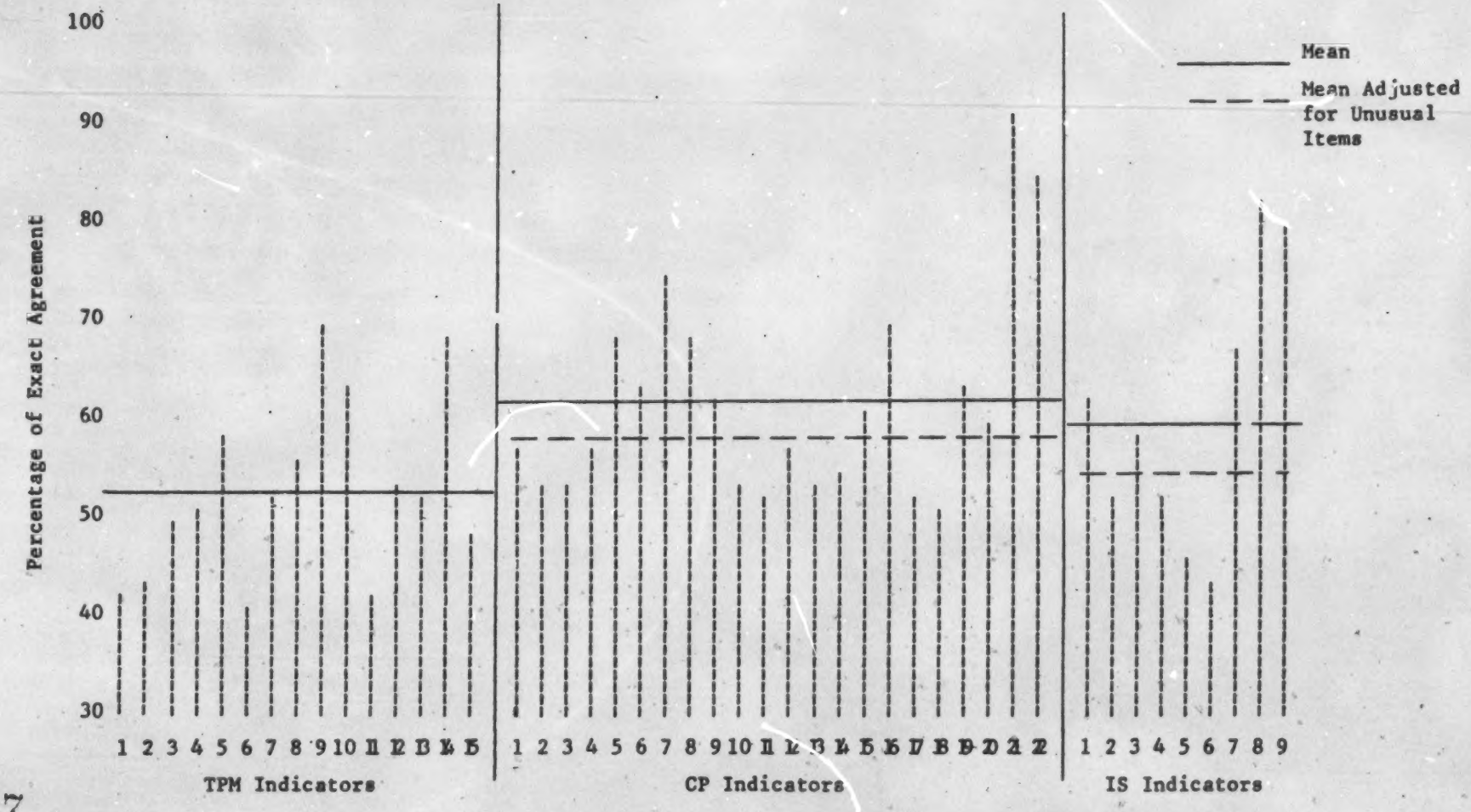
The IS instrument appears to have three distinct levels of agreement among its indicators. Indicators 1 through 4 have means similar to the instrument mean (.60). Indicators 5 and 6 have relatively low rates of agreement; and the balance seem to be somewhat easier for groups to rate consistently. The last few indicators assess situations which do not occur frequently on the videotapes that have been used in the training program. Therefore, the range of plausible choices is limited for the raters.

The findings of this study are similar to results generated in previous studies. It seems consistent with earlier suggestions that agreement is more difficult to achieve when videotaped or other standardized situations are used than when "live" teaching performances are rated. The lower rates of agreement for the TPM may be explained by the difficulty in having the trainers maintain a standard role as models throughout the large number of interviews that were conducted.

As in prior years' work, Near Agreement rates were considerably higher than Exact Agreement. This suggests that whenever disagreements occured among raters they were generally minor ones.

A final measure of agreement is the congruence between a panel, that provided the "right" answers used in training, and the much larger group of trainees. The difference in the mean ratings of the two groups in these lessons was .30 on the TPM and .25 on the CP and IS.

Figure 1


Mean
Mean Adjusted for Unusual Items
Percentage of Exact Agreement
100
90
80
70
60
50
40
30
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
TPM Indicators
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
CP Indicators
1 2 3 4 5 6 7 8 9
IS Indicators

Rates of Agreement on TPAI Indicators

Internal Consistency of Ratings using the TPAI (Beginning Teacher Edition)

The Problem

When a number of trained observers rate a professional teacher's performance on the TPAI (via videotape) or the products (lesson plan portfolio), to what extent are their ratings internally consistent?

Procedures

Data for these analyses were collected during the 1978 summer training sessions (for leadership personnel) conducted in Athens. Six training sessions were conducted. Participants in these sessions held a wide variety of experience and educational positions. Included were college professors, teachers and other educational specialists from school districts, CESA's, and the State Department of Education.

During the course of the workshop activity, participants viewed videotapes to rate several lessons using the Classroom Procedures (CP) and the Interpersonal Skills (IS) Assessment Instruments. They also evaluated sets of lesson plans using the Teaching Plans and Materials (TPM) Assessment Instrument.

For each session where a sufficient number of trainees rated the same tape or lesson plan portfolio, Cronbach's Alpha was calculated for the ratings. Thus, there were as many as three replications of the studies in some cases. In addition, data for each tape or portfolio were combined for all sessions so that a more powerful analysis could be made.

A second analysis was made of the data combined from the various training sessions. Groups of three individuals were randomly drawn from the raters and their ratings averaged for each indicator. These averages were then used as the basis for calculating Alpha coefficients. This

combination of ratings is what would occur if three observers all rated a teaching performance based on the same visit to the classroom, a procedure similar to that planned for implementation.

Results

Table five presents Cronbach Alpha coefficients of individual raters for two TPM portfolios. Alpha values ranged from .64 to .81 on nine occasions. The Alpha estimates for combined groups of raters across all rating occasions were .73 (Third Grade Science) and .76 (Eighth Grade Language Arts).

The Alpha values raters for the CP and IS instruments for four videotapes are shown in Table 6. For the CP the range was .58 to .86 and for the IS the range was .50 to .77. Combined group Alpha values were .88, .80, .79 and .79 for the CP, and .79, .71, .51 and .61 for the IS.

When sets of three individuals were randomly selected from the combined groups and the mean of the three ratings used to construct an Alpha coefficient the values were higher than with single raters in nearly every case. For the TPM the Alpha values were .74 and .[illegible] (See Table 5). For the CP the values were .94, .90, .57 and .60. For the IS the Alpha values were .87, .77, .60, and .33. The lowest estimates for both the CP and IS were associated with the same videotape--The Ninth Grade Consumer Economics Teacher.

Table 5

Cronbach Alpha Estimates of the Reliability of the Teaching Plans and Materials Instrument*

Lesson Plan Portfolio	Number of Raters**	Alpha Value
Third Grade Science		
First Occasion	15	.78
Second Occasion	15	.68
Third Occasion	20	.65
Fourth Occasion	20	.74
Fifth Occasion	12	.69
Sixth Occasion	19	.77
Combined Group	99	.73
Sets of Three Raters	25 X 3 = 75	.74
Eighth Grade Language Arts		
First Occasion	24	.64
Second Occasion	14	.81
Third Occasion	18	.81
Combined Group	64	.76
Sets of Three Raters	19 X 3 = 57	.79

* The instrument has 15 indicators

**The numbers may not total over all occasions to equal the combined group since the analyses were completed at different times. Also each set of three raters was drawn from the same rating occasion. Therefore if the number of raters was not a multiple of three some of the raters were lost.

Table 6

Cronbach Alpha Estimates for the Classroom Procedures and Interpersonal Skills Instruments*

Videotape Lesson	Classroom Procedures Number of Raters**	Classroom Procedures Alpha Value	Interpersonal Skills Number of Raters**	Interpersonal Skills Alpha Value
Third Grade Science				
First Occasion	16	.75	8	.49
Second Occasion	13	.62	10	.58
Third Occasion	17	.71	6	.65
Fourth Occasion	20	.83	16	.77
Combined Group	66	.88	45	.79
Sets of Threes	20 X 3 = 60	.94	22 X 3 = 66	.87
Eighth Grade Language Arts				
First Occasion	24	.67	18	.61
Second Occasion	17	.84	11	.50
Third Occasion	9	.74	6	.65
Fourth Occasion	18	.58		
Combined Group	68	.80	48	.71
Sets of Threes	22 X 3 = 66	.90	23 X 3 = 69	.77

Table 6 continued.

Videotaped Lesson	Classroom Procedures Number of Raters**	Classroom Procedures Alpha Value	Interpersonal Skills Number of Raters**	Interpersonal Skills Alpha Value
Intermediate Language Arts				
First Occastion	7	.68	16	.75
Combined Group	14	.51	29	.79
Sets of Threes	9 X 3 = 27	.57	9 X 3 = 27	.87
Ninth Grade Consumer Economics				
First Occasion	14	.71	5	.72
Second Occasion	11	.86	5	.50
Third Occasion	16	.79	9	.69
Combined Group	41	.79	19	.61
Sets of Threes	12 X 3 = 36	.60	12 X 3 = 36	.33

* The Classroom Procedures has 22 indicators; the Interpersonal Skills has 9 indicators.

** See botton of Table Five.

Discussion

The Alpha coefficients of greatest concern are those based on the means of three observers, since their situation is closest to actual TPAI administration procedures. The internal consistency of the TPM, CP, and IS instruments estimated in this way is acceptable in most cases. The exception is the Ninth Grade Consumer Economics tape. This particular tape was rated by twelve groups of three raters each for a total of 36 individual ratings. It may be that when the groups of three were constituted as the data were assembled a chance combination of groups produced low ratings in this instance. Additional analyses of these data are being planned to assess this possibility. Another caution regarding this particular tape is that it was used somewhat earlier in the training activities than any of the others for which data are available. Consequently, there may be a problem in the ratings themselves.

When these results are compared with analagous results of earlier studies (Capie, et al, 1978), a slight drop in the internal consistency of the ratings is noted--.76 to .74 on the TPM; .91 to .82 on the CP; and .73 to .66 on the IS. Such a finding initially disappoints; but there has been a large increase in interrater agreement from 1978 to 1979. Consequently, the variance of ratings has been decreased and Cronbach's Alpha decreased as well. The large increases in interrater agreement reported from 1978 to 1979 compensate for the slight decreases in the Alpha reliabilities.

Cronbach's Alpha, in this case, might be judged as supporting both the validity and the reliability of the TPAI. From this viewpoint: it seems that the IS indicators are less homogeneous than indicators for

the other two instruments. It may be that the IS instrument measures two or more constructs as is suggested by the factor analytical study reported by Tobin et al. (1979). The IS appears to have two components, one which assesses classroom management and another which assesses enthusiasm and sensitivity to students.

The TPM and CP, in particular, seem to have high levels of construct validity as estimated by the Alpha coefficients.

Interrater Agreement Using the TPAI (Beginning Teacher Edition)

The Problem

When two or more trained data collectors observe a professional performance, to what extent do they use the same descriptor rating scale point to describe that teaching performance?

Procedures

The study was conducted in the Griffin CESA district at the conclusion of the classroom portion of the data collector training program.

A total of 72 obseravtions were made in conducting assessments of 30 teachers. Three or four observers were involved in each assessment, they completed independent ratings after simultaneous interview and classroom observation.

Each rater completed the Teacher Plans and Materials (TPM), Classroom Procedures (CP), and Interpersonal Skills (IS) instruments.

Two similar types of data analyses were completed. Each resulted in an index (proportion) of agreement among the raters. The first index was Exact Agreement, which amounted to choosing the same or identical rating for an indicator. In other words, the portion of raters choosing the mode was determined. The second index was Near Agreement, which amounted to selecting either the mode or the next most common adjacent scale point. For each teacher rated, the ratings on each indicator were compared across all raters to determine the proportion of agreement. The mean levels of exact and near agreement were determined for each indicator. Examples of these procedures are shown in Table 7. Instances in which raters failed to rate an indicator were treated as numerical ratings when computing exact-agreement. Thus, if all raters

agreed that an indicator was not rateable 100% exact and near agreement were noted. Instances where some rater failed or abstained to rate an indicator were recorded as "misses" when Near Agreement was recorded.

Table 7

Sample Ratings and Proportion of Agreement

Teacher	Indicator	Individual Ratings A	B	C	D	Proportion of Exact Agreement	Proportion of Near Agreement
A	CP 1	4	4	3	3	2/4 = .50	4/4 = 1.00
B	TPM 5	4	4	3	2	2/4 = .50	3/4 = .75
C	IS 3	5	4	4	3	2/4 = .50	3/4 = .75
D	TPM 11	5	4	3		1/3 = .00	2/3 = .67
E	CP 12	5	4	4		2/3 = .67	3/3 = 1.00
F	CP 22	-	4	4		2/3 = .67	2/3 = .67
Mean Proportions						1/2 = .50	18/21 = .86

Results

The percentages of near and exact Agreement for each indicator of the Teaching Plans and Materials (TPM) instrument are shown in Table 8. The mean levels of Exact Agreement ranged from .77 (Indicator 2) to .92 (Indicator 10). Near Agreement ranged from .85 to 1.00 (Indicator 2 and 7), respectively. The mean near and exact agreement rates for the TPM wer .86 and .98, respectively.

The interrater agreement results for the Classroom Procedures (CP) instrument are shown in Table 9. Exact Agreement ranged from .81 (indicators 13 and 17) to .95 (indicator 7). Near Agreement ranged from .92 (indicator 17) to 1.00 (indicators 2, 3, 7, 8, and 19). The means for these two indices were .87 and .98.

Table 10 summarizes interrater agreement data for the Interpersonal Skills (IS) instrument. Exact agreement ranged from .83 (Indicator 5 and 6) to .98 (Indicator 9). Near agreement ranged from .94 (Indicator 6) to 1.00 (Indicators 7, 8, and 9). Means of .90 and .98 were calculated for the two indices.

Table 8
Interrater Agreement Using the
Teaching Plans and Materials (TPM)
with 30 Teachers

Indicator	Exact Agreement	Near Agreement
1	.83	.96
2	.77	.85
3	.89	.98
4	.88	.98
5	.83	.95
6	.85	.98
7	.87	1.00
8	.88	.99
9	.89	.96
10	.92	.99
11	.82	.96
12	.86	.95
13	.82	.96
14	.89	.96
15	.87	.99
Mean	.86	.98

Table 9
Interrater Agreement Using the
Classroom Procedures (CP)
with 30 Classrooms

Indicator	Exact Agreement	Near Agreement
1	.92	.98
2	.88	1.00
3	.87	1.00
4	.90	.98
5	.88	.98
6	.91	.98
7	.95	1.00
8	.86	1.00
9	.86	.99
10	.84	.97
11	.90	.98
12	.85	.96
13	.81	.94
14	.88	.98
15	.91	.99
16	.89	.97
17	.81	.92
18	.85	.95
19	.87	1.00
20	.84	.96
21	.86	.94
22	.91	.99
Mean	.87	.98

Table 10

Interrater Agreement Using the Interpersonal Skill (IS) with 30 Classroom

Indicator	Exact Agreement	Near Agreement
1	.91	.98
2	.86	.95
3	.92	.97
4	.87	.97
5	.83	.98
6	.83	.94
7	.94	1.00
8	.97	1.00
9	.98	1.00
Mean	.90	.98

Discussion

The most notable observation about these data is the remarkable improvement in interrater agreement over previously reported findings. In 1978, Near Agreement rates were approximately 70 percent (Capie, et. al). Averages 20 to 30 percent higher resulted from this year's study. Such levels of agreement are satisfactory for an instrument with the moderate inference level of the TPAI.

As with the 1978 study, trainees were used as teachers when establishing interrater agreement. While such a special population might suggest that agreement is attributable to a limited range of plausible ratings, the data do not support this theory. Table 11 shows the proprotion of occasions on which each descriptor was chosen (a total of 3322 ratings or 72 raters X46 indicators). While high ratings predominated, low ratings did occur. Chance probably does not account for the increase in agreement from 1978 to 1979.

A second observation is that the rates of agreement are considerably higher in the classroom as compared with the videotape studies reported earlier (approximately 65% as compared with 85%). It was interesting to note that rates of agreement among groups of 3 to 4 raters correlated strongly whenever ratings of videotaped lessons were compared with ratings based on in-the-classroom observation.

The rates of agreement in this study were uniformly high. As shown in Figure 2, there were few indicators with unusually low rates of agreement.

The rates of agreement will probably not improve greatly through subsequent revisions of instruments on training processes. It is encouraging to not that coefficients of internal consistency for the TPAI

instruments have not dropped greatly in spite of the low variation among ratings attributable to high interrater agreement.

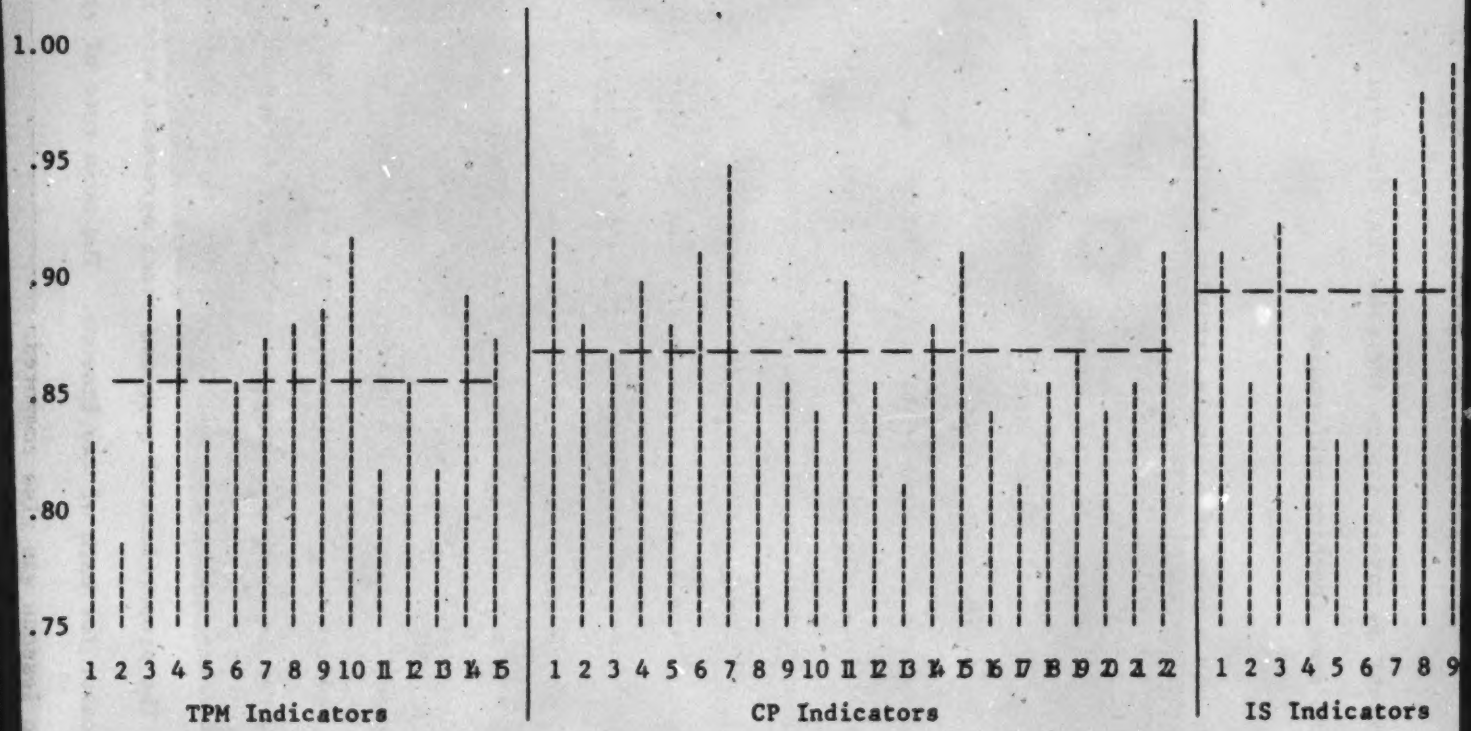
The increases that have been noted in the interrater agreement were anticipated based on the development and training activities that occurred in the Spring and Summer of 1978. Indicators and descriptors were revised based on earlier study and anecdotal feedback. Similarly, the training program was improved considerably so that the variation in the interpretations of the instruments was reduced. Finally, some of these individuals have a great deal of experience in using the instruments so that experience may have accounted for some of the improvement.

Table 11

Proportion of Scale Point Selection Reported as Percent of One Hundred

Scale of Descriptors					
1	2	3	4	5	X
1.5	5.8	20.8	30.1	36.3	5.5

Figure 2



Rates of Agreement on TPAI Indicators
in Classroom Assessment

Interrater Agreement Using the TPAI in Training Situations (Student Teacher Edition)

The Problem

To what extent do raters assign the same TPAI descriptor rating when they observe the same teaching performance?

Procedures

This study was a part of a single training session conducted in November of 1978 for college supervisors of Student teachers. A total of 26 professors from Georgia colleges of Education participated.

The Teaching Plans and Materials (TPM) data were collected as part of the evaluation of the trainees near the end of their training. Each professor examined one of two portfolios and then met with a trainer who played a standard role adapted to the student teacher version of the instruments. The ratings were based on the assessment of th plans and the information obtained in the interview. The trainees then watched a videotape of one of the lessons in the portfolio. At the conclusion of this lesson the Classroom Procedures (CP) and Interpersonal Skills (IS) indicators were rated.

Two similar types of data analyses were completed. Each resulted in an index (proportion) of agreement among the raters. The first index was Exact Agreement, which amounted to choosing the same or identical rating for an indicator. In other words, the portion of raters choosing the mode was determined. The second index was Near Agreement, which amounted to selecting either the mode or the next most common adjacent scale point. The proportions of near and exact agreement were calculated for each indicator for each of two lessons. The mean rate of agreement across the two lessons was also computed.

Instances in which all or most of the raters failed or abstained to rate an item were still treated as numerical ratings when determining Exact Agreement. Thus, if all raters agreed that an indicator was not rateable then 100 percent Exact and Near agreement were noted. Instances where some raters failed or abstained to rate an item were recorded as "misses" when Near Agreement was computed. Samples of these procedures are shown in Table 12.

Table 12

Sample Ratings and Proportion of Agreement

Teacher	Indicator	Number of Ratings for each Rating Point						Proportion of Exact-Agreement	Proportion of Near Agreement
		1	2	3	4	5	X		
A	CP 1			14	13			.52	1.00
B	TPM 5			4	13	12		.45	.86
C	IS 3			14	14			.50	.50
D	TPM 11			15	3		9	.56	.67
E	CP 22				3		24	.89	.89

Results

Table 13 displays the results of the two analyses of the TPM Indicator portfolios. The mean rate of Exact Agreement varied between .39 and .64 (indicators 7 and 5, respectively). The mean level of Exact Agreement across the 12 TPM indicators was .52. Near Agreement ranged from .61 (indicator 12) to .93 (indicators 3 and 5). The mean of Near Agreement was .80 for the indicators. Mean rates of exact and near agreement for the two sets of lesson plans were almost identical.

Table 14 presents the near and exact agreement that resulted after the use of the CP instrument as applied to two videotaped lessons. The mean rate of Exact Agreement for the CP was higher than for the TPM - .64 across the 22 indicators. The indicator means ranged from .43 (indicator 1) to 1.00 (indicators 21 and 22). The mean rate of Near Agreement for the CP indicators for the two lessons combined was .90.

The results of the analyses of the IS instrument for the two lessons are shown in Table 15. The mean level of Exact Agreement spread from .50 (indicator 6) to .96 (indicators 8 and 9). The mean rate of Exact Agreement across the nine indicators was .66. The mean rate of Near Agreement for the IS was .90. Indicator means ranged from .67 (indicator 5) to 1.00 (indicators 4, 9 and 10).

Table 13

Interrater Agreement using the TPM
for Student Teacher Portfolios
with Standard Interviews

Indicator	8th Grade Language Arts n=12		3rd Grade Science n=16		Weighted Mean n=26	
	Exact Agreement	Near Agreement	Exact Agreement	Near Agreement	Exact Agreement	Near Agreement
1	.42	.83	.44	.81	.43	.82
2	.50	.67	.50	.94	.50	.82
3	.42	.83	.63	1.00	.54	.93
4	.58	.83	.50	.69	.53	.75
5	.58	.92	.69	.94	.64	.93
6	.58	.75	.38	.69	.47	.72
7	.50	.92	.31	.50	.39	.68
8	.67	.92	.56	.88	.51	.90
9	.50	.83	.69	.88	.61	.86
10	.75	.92	.44	.69	.57	.79
11	.50	.67	.69	.81	.61	.75
12	.33	.50	.56	.69	.46	.61
Mean*	.53	.80	.53	.79	.52	.80

*Column and row means are altered by rounding differences.

Table 14

Interrater Agreement Using the CP with Videotaped Lessons of Student Teachers

Indicator	8th Grade Language Arts n=11		3rd Grade Science n=13		Weighted Mean n=24	
	Exact	Near	Exact	Near	Exact	Near
1	.42	.82	.38	.54	.43	.69
2	.64	.91	.69	.92	.67	.92
3	.64	1.00	.69	.92	.67	.96
4	.98	.82	.46	.77	.47	.79
5	.55	.91	.62	.85	.59	.88
6	.82	1.00	.54	1.00	.67	1.00
7	.82	.91	.62	.92	.71	.92
8	.64	.91	.54	1.00	.59	.96
9	.45	.91	.54	.77	.50	.83
10	.36	.73	.62	.85	.50	.79
11	.64	1.00	.77	1.00	.71	1.00
12	.73	1.00	.69	.92	.71	.96
13	.36	.64	.62	.92	.50	.79
14	.36	.64	.69	1.00	.54	.84
15	.55	.73	.69	.77	.63	.75
16	.64	.91	.46	.92	.54	.92
17	.73	.91	.46	.92	.58	.92
18	.64	.91	.69	.85	.67	.88
19	.64	.91	.92	1.00	.79	.96
20	.64	.91	.62	1.00	.63	.96
21	1.00	1.00	1.00	1.00	1.00	1.00
22	1.00	1.00	1.00	1.00	1.00	1.00
Mean	.63	.89	.65	.86	.64	.90

Table 15

Interrater Agreement Using the IS
With Videotaped Lessons-
Student Teachers

Indicator	8th Grade Language Arts (n=11) Exact	Near	3rd Grade Science (n=15) Exact	Near	Weighted Mean (n=26) Exact	Near
1	.64	.91	.73	.93	.69	.92
2	.55	.91	.47	.93	.50	.92
3	.55	.91	.60	.93	.58	.92
4	.55	1.00	.60	1.00	.58	1.00
5	.38	.75	.60	.60	.51	.67
6	1.36	.64	.60	.80	.50	.73
7	.55	1.00	.73	.93	.65	.96
8	1.00	1.00	.93	1.00	.96	1.00
9	1.00	1.07	.93	1.00	.96	1.00
Mean	.62	.90	.69	.90	.66	.90

Discussion

These rates of agreement are similar to those that are reported for the Beginning Teacher Edition of the instruments. That is not surprising since many of the same items are common to both and the same videotaped stimuli were used in both training sessions and both studies. The agreement rates are slightly higher with the college supervisors reported in this study than with the public school personnel involved in other studies of the Beginning Teacher Instrument. That slight difference favoring the college supervisors is somewhat unusual in light of the large difference in agreement rates when these same individuals rated a real classroom setting. The college supervisors were actually able to agree about the quality of a videotaped lesson as easily as a lesson viewed in the classroom. Public school personnel, on the other hand, achieved large increases in agreement when they worked in the classroom as opposed to videotape.

A similar pattern of problems emerged in this study. Certain indicators, such as IS5 and IS6, were characterized by low levels of near- as well as exact agreement. These indicators are among those which received great scrutiny during the current (1979-1980) revision of the instruments. Some indicators had low levels of agreement in one of the two situations. For example, the agreement rate of TPM 7 in the Third Grade Science was lower than for the Language Arts which suggests that the ambiguity may rest with the standard interview rather than with the instrument.

These results do suggest that the status of the teacher being assessed does not seem to be related to the levels of agreement among the raters. Similarly, the mean ratings were nearly identical in both teaching contexts. Such a finding would suggest that using the TPAI does aid the objectivity of the observer.

Internal Consistency of Ratings using the TPAI (Student Teacher Edition)

The Problem

When a number of trained observers rate a student teacher's performance on the TPAI (via videotape) or the products (lesson plan portfolio), to what extent are their ratings internally consistent?

Procedures

Data for these analyses were collected during a training session conducted in November, 1978 for college supervisors of student teachers. A total of 26 professors from Georgia colleges participated in the session.

The Teaching Plans and Materials instrument data were collected as part of the evaluation of the trainees. The trainee examined a portfolio and then was scheduled for an interview with a trainers who played a standard role adapted for the student teacher version. The ratings were based on the assessment of the plans and the information included in the interview.

Each trainee then viewed a videotaped lesson in which one of the lessons in the portfolio was implemented. At the conclusion of the lesson the Classroom Procedures (CP) and the Interpersonal Skills (IS) were rated. The number of available ratings for the TPM, CP, and IS varied from a low of 11 to a high of 16.

Cronbach Alpha internal consistency coefficients were computed for the ratings on each of the three instruments for two portfolio and lesson assessments.

Results

The results of the analyses are listed in Table 16. The Alpha values were 0.80 and 0.75 for the TPM, 0.88 and 0.75 for the CP, and 0.65 and 0.44 for the IS.

Table 16

Cronbach Alpha Values
for the Student Teacher Version of the TPAI

Instrument	3rd Grade Science		8th Grade Language Arts	
	Number of Raters	Alpha	Number	Alpha
Teaching Plans & Materials	.80	16	.75	12
Classroom Procedures	.75	13	.88	11
Interpersonal Skills	.65	15	.44	11

Discussion

These same materials have been used with many other groups in association with the beginning teacher version of the instruments. The Alpha values generated for the TPM and CP using the student teacher supervisors' ratings were similar to other reported findings. The IS Alpha coefficients, however are somewhat lower than the previous findings for both of these tapes. The variation within ratings for the IS was less than that for the other two instruments. This fact may account for the relatively low Alpha reliabilities.

Interrater Agreement Using the TPAI

(Student Teacher Edition)

Problem

When three or more trained raters observe a teaching performance, to what extent do they use the same descriptor ratings scale point to describe a teaching performance?

Procedure

Raters for this study were 24 college supervisors of student teachers who were experiencing a field practice activity in conjunction with a training session conducted in November, 1978. These individuals represented a number of colleges throughout Georgia.

"Student" teachers for this study were 6 University of Georgia early childhood education majors who were enrolled in a pre-student teaching experience. The actual status of these students was not disclosed to the raters. Plans for the day's lesson were made available to raters, three or four of whom visited each classroom. After a period of approximately 60 minutes, each observer independently completed the Classroom Procedures (CP) and Interpersonal Skills (IS) instruments. Discussion of ratings was then used to complete the training activity.

For each student teacher rated the ratings on each indicator were compared across raters. Two similar types of analyes were completed. Each resulted in an index (proportion) of agreement among raters. The first index was Exact-Agreement, choosing the same or identical rating for an indicator. In other words, the number of persons choosing the modal response was noted. The second index was Near-Agreement, choosing the mode or the next most common, adjacent scale point. The mean levels

Discussion

These same materials have been used with many other groups in association with the beginning teacher version of the instruments. The Alpha values generated for the TPM and CP using the student teacher supervisors' ratings were similar to other reported findings. The IS Alpha coefficients, however are somewhat lower than the previous findings for both of these tapes. The variation within ratings for the IS was less than that for the other two instruments. This fact may account for the relatively low Alpha reliabilities.

Interrater Agreement Using the TPAI

(Student Teacher Edition)

Problem

When three or more trained raters observe a teaching performance, to what extent do they use the same descriptor ratings scale point to describe a teaching performance?

Procedure

Raters for this study were 24 college supervisors of student teachers who were experiencing a field practice activity in conjunction with a training session conducted in November, 1978. These individuals represented a number of colleges throughout Georgia.

"Student" teachers for this study were 6 University of Georgia early childhood education majors who were enrolled in a pre-student teaching experience. The actual status of these students was not disclosed to the raters. Plans for the day's lesson were made available to raters, three or four of whom visited each classroom. After a period of approximately 60 minutes, each observer independently completed the Classroom Procedures (CP) and Interpersonal Skills (IS) instruments. Discussion of ratings was then used to complete the training activity.

For each student teacher rated the ratings on each indicator were compared across raters. Two similar types of analyes were completed. Each resulted in an index (proportion) of agreement among raters. The first index was Exact-Agreement, choosing the same or identical rating for an indicator. In other words, the number of persons choosing the modal response was noted. The second index was Near-Agreement, choosing the mode or the next most common, adjacent scale point. The mean levels

of exact and near agreement were computed for each indicator acorss the six classrooms. Examples of these procedures are shown in Table 17. Instances in which raters failed or abstained to rate an item were treated as numerical ratings when computing Exact Agreement. Thus, if all raters agreed that an indicator should not be rated 100% exact and near agreement were noted. Instances where one rater considered an item not rateable were treated as "misses" when Near Agreement was computed.

Table 17
Sample Ratings and Proportion of Agreement

Teacher	Indicator	Individual Ratings A B C D	Proportion of Exact-Agreement	Proportion of Near-Agreement
A	CP 1	4 4 3 3	2/4 = .50	4/4 = 1.00
B	TPM 5	4 4 3 2	2/4 = .50	3/4 = .75
C	IS 3	5 4 4 3	2/4 = .50	3/4 = .75
D	TPM 11	5 4 3	1/4 = .33	2/3 = .67
E	CP 27	5 4 4	2/3 = .67	3/3 = 1.00
F	CP 17	- 4 4	2/3 = .67	2/3 = .67
Mean Proportions			10/22 = .45	18/21 = .86

Results

The mean levels of exact and near agreement for the CP and IS instruments are shown in Tables 18 and 19. For the CP, Exact Agreement ranged from 0.54 (Indicator 1) to 0.96 (Indicator 9); Near Agreement ranged from 0.62 (Indicatorl) to 1.00 (Indicator 9). The mean rate of Exact Agreement was 0.69 and Near Agreement 0.88.

For the IS, Exact Agreement ranged from 0.54 to 0.96 (Indicator 5 and 4, respectively); Near Agreement ranged from 0.67 (Indicator 5) to 1.00 (Indicator 5, 4, 6, and 7). The mean rate of Exact Agreement was 0.73.

Table 18
Interrater Agreement Using the Classroom Procedures Instrument with Student Teachers in Six Classrooms

	Mean Level of Exact Agreement	Mean Level of Near Agreement
1	.54	.62
2	.67	.96
3	.63	.87
4	.58	.87
5	.58	.87
6	.71	.96
7	.58	.87
8	.67	.87
9	.96	1.00
10	.62	.92
11	.67	.96
12	.92	.96
13	.62	.79
14	.62	.92
15	.75	.92
16	.83	.96
17	.63	.79
18	.58	.92
19	.63	.87
20	.67	.83
21	.87	.87
22	.83	.83
Mean	.69	.88

Table 19
Interrater Agreement Using the Interpersonal Skills Instrument with Student Teachers

Indicator	Exact Agreement	Near Agreement
1	.58	.92
2	.62	.75
3	.79	.79
4	.96	1.00
5	.54	.67
6	.75	1.00
7	.92	1.00
8	.75	.87
9	.67	.92
Mean	.73	.88

Discussion

These results do not show agreement levels as high as those for samples of beginning teachers. Several reasons might account for these findings. No TPM was completed, so the context of the lesson was unknown. Furthermore, unlike studies of beginning teachers, none of these raters was familiar with the teaching situation or the school system. A second possible problem is the fact that many of these supervisors were not familiar with the early childhood education context, which may have led to more diverse ratings. A third difference is that one of the beginning teacher observers was a "trainer" whereas all of the present observers were using the student teacher instrument for the first time. A fourth concern is the observation that college supervisors tend to be more unreliable in their ratings than do other trainees during the course of the training.

Nevertheless, mean rates of agreement are higher in this study than those obtained in 1978 with beginning teachers in the classroom (Capie, et al). And, experience does seem to be a factor in the rise of rates of agreement. Several indicators on which agreement rates were low have been revised for the 1979 edition of the instruments--namely: 1, 4, 5, and 18; IS 1 and 5.

Although Near Agreement rates are somewhat lower than in studies using beginning teachers, an improvement in agreement rates becomes obvious when Near Agreement is considered. This finding suggests that when disagreement occurs it is consistent in direction.